

Bayesian density regression for count data

Charalampos Chaniailidis¹, Ludger Evers², and Tereza Neocleous³

¹University of Glasgow, c.chaniailidis1@research.gla.ac.uk

²University of Glasgow, ludger.evers@glasgow.ac.uk

³University of Glasgow, tereza.neocleous@glasgow.ac.uk

June 10, 2014

Abstract

Despite the increasing popularity of quantile regression models for continuous responses, models for count data have so far received little attention. The main quantile regression technique for count data involves adding uniform random noise or “jittering”, thus overcoming the problem that the conditional quantile function is not a continuous function of the parameters of interest. Although jittering allows estimating the conditional quantiles, it has the drawback that, for small values of the response variable Y , the added noise can have a large influence on the estimated quantiles. In addition, quantile regression can lead to “crossing” quantiles. We propose a Bayesian Dirichlet process (DP)-based approach to quantile regression for count data. The approach is based on an adaptive DP mixture (DPM) of COM-Poisson regression models and determines the quantiles by estimating the density of the data, thus eliminating all the aforementioned problems. Taking advantage of the exchange algorithm, the proposed MCMC algorithm can be applied to distributions on which the likelihood can only be computed up to a normalising constant.

1 Quantile regression

Quantile regression was introduced as a nonparametric method for modelling a variable of interest as a function of covariates [6]. By estimating the conditional quantiles rather than the mean, it gives a more complete description of the conditional distribution of the response variable than least squares regression, and is especially relevant in certain types of applications.

Consider a random variable Y with cumulative distribution function $F(y)$. The p th quantile function of Y is defined as

$$Q(p) = \inf\{y \in \mathbb{R} : p \leq F(y)\} \quad (1)$$

and can be obtained by minimising the expected loss $E[\rho_p(Y - u)]$ with respect to u , where $\rho_p(y) = |y(p - I(y < 0))|$. The p th sample quantile is obtained in a similar way by minimising $\sum_{i=1}^n \rho_p(y_i - u)$.

Suppose that the p th conditional quantile function, $Q_Y(p|X = \mathbf{x})$, is a linear function of the predictors so that $Q_Y(p|X = \mathbf{x}) = X'\beta_p$. The parameter estimates $\hat{\beta}_p$ are then obtained as

$$\hat{\beta}_p = \arg \min_{\beta_p \in \mathbb{R}^k} \sum_{i=1}^n \rho_p(Y - X'\beta_p). \quad (2)$$

A closed-form solution for this minimisation problem does not exist since the objective function is not differentiable at the origin, and it is solved using linear programming techniques [1].

1.1 Quantile regression for count data

The problem with applying quantile regression to count data is that the cumulative distribution function of the response variable is not continuous, resulting in quantiles that are not continuous, and which thus can not be expressed as a continuous function of the covariates. One way to overcome this problem is by adding uniform random noise (“jittering”) to the counts [7]. The general idea is to construct a continuous variable whose conditional quantiles have a one-to-one relationship with the conditional quantiles of the counts. Defining the new continuous variable $Z = Y + U$ where Y is the count variable and U is a uniform random variable in the interval $[0, 1)$, the conditional quantiles $Q_Z(p|X = \mathbf{x}) = p + \exp(X'\beta_p)$.

The variable Z is transformed in such a way that the new quantile function is linear in the parameters, i.e. $Q_{T(Z;p)}(p|X = \mathbf{x}) = X'\beta_p$ where

$$T(Z; p) = \begin{cases} \log(Z - p) & \text{for } Z > p, \\ \log(\varsigma) & \text{for } Z \leq p, \end{cases} \quad (3)$$

with ς being a small positive number. The parameters β_p are estimated by running a linear quantile regression of $T(Z; p)$ on x . Finally, the conditional quantiles of interest, $Q_Y(p|X = \mathbf{x})$ can be obtained from the previous

quantiles as

$$Q_Y(p|X = \mathbf{x}) = \lceil Q_Z(p|X = \mathbf{x}) - 1 \rceil \quad (4)$$

where $\lceil p \rceil$ denotes the ceiling function which returns the smallest integer greater than, or equal to, p .

While the jittering approach eliminates the problem of a discrete response distribution, for small values of the response variable Y , the mean and the variance in the transformed variable Z will be mainly due to the added noise, resulting in poor estimates of the conditional quantiles $Q_Y(p|X = \mathbf{x})$. As an example, when $Y = 0$ the term $\log(Z - p) = \log(U - p)$ could go from $-\infty$ to 0, simply due to the added noise. In addition, quantile regression can suffer from the problem of crossing quantile curves, which is usually seen in sparse regions of the covariate space. This happens due to the fact that the conditional quantile curve for a given $X = \mathbf{x}$ will not be a monotonically increasing function of p .

Another approach would be to view the counts as ordinal variables with fixed thresholds and then model the new latent variable by an infinite mixture of normal densities [5]. Instead of using the aforementioned methods, we propose an adaptive Dirichlet process mixture approach which estimates the conditional density of the data. The approach is based on an adaptive Dirichlet Process mixture (DPM) of COM-Poisson regression models.

2 COM-Poisson distribution

The COM-Poisson distribution [2, 11] is a two-parameter generalisation of the Poisson distribution that allows for different levels of dispersion. The probability mass function of the COM-Poisson(λ, ν) distribution is

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)} \quad y = 0, 1, 2, \dots \quad (5)$$

where $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ and $\lambda > 0$ and $\nu \geq 0$, where the normalisation constant does not have a closed form and has to be approximated numerically. The extra parameter ν allows the distribution to model under- ($\nu > 1$) or over-dispersed ($\nu < 1$) data, having the Poisson distribution as a special case ($\nu = 1$).

The above formulation of the COM-Poisson does not have a clear centering parameter since the parameter λ is close to the mean only when ν takes values close to 1, which makes it difficult to interpret for under- or

over-dispersed data. Substituting the parameter λ with $\mu = \lambda^{\frac{1}{\nu}}$, where $\lfloor \mu \rfloor$ is the mode of the distribution

$$\mathbb{E}[Y] \approx \mu, \quad \mathbb{V}[Y] \approx \frac{\mu}{\nu} \quad (6)$$

and the new probability mass function is

$$P(Y = y | \mu, \nu) = \left(\frac{\mu^y}{y!} \right)^\nu \frac{1}{Z(\mu, \nu)} \quad y = 0, 1, 2, \dots \quad (7)$$

where $Z(\mu, \nu) = \sum_{j=0}^{\infty} \left(\frac{\mu^j}{j!} \right)^\nu$.

2.1 Mixtures of COM-Poisson distributions

The COM-Poisson is flexible enough to approximate distributions with any kind of dispersion in contrast to a Poisson or a mixture of Poisson distributions which can only deal with overdispersion.

The two parameters of the COM-Poisson distribution allow it to have arbitrary (positive) mean and variance; one can obtain a point mass by letting the variance parameter ν tend to infinity. Thus one can show that mixtures of COM-Poisson distributions can provide an arbitrarily precise approximation to any discrete distribution with support \mathbb{N}_0 , which is why COM-Poisson distributions are used by our method. All other generalisations of the Poisson distribution we are aware of do not have this property.

2.2 COM-Poisson regression

A regression model can be defined based on (7), in which both the mean and the variance parameter are modelled as a function of covariates:

$$\log \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta} \quad (8)$$

$$\log \nu_i = \mathbf{x}_i^\top \mathbf{c} \quad (9)$$

where Y is the response variable being modelled, and $\boldsymbol{\beta}, \mathbf{c}$ are the regression coefficients for the centering link function and the shape link function respectively. The parameters in this formulation have a direct link to either the mean or the variance, providing insight into the behaviour of the response variable. Notably,

$$\mathbb{E}[Y_i] \approx \exp(\mathbf{x}_i' \boldsymbol{\beta}), \quad \mathbb{V}[Y] \approx \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{\exp(\mathbf{x}_i' \mathbf{c})} = \exp(\mathbf{x}_i' (\boldsymbol{\beta} - \mathbf{c})). \quad (10)$$

The calculation of the normalisation constant of the COM-Poisson distribution is the computationally most expensive part of the proposed regression model. It can be seen, in the next subsection, that this calculation is redundant.

2.3 Exchange algorithm

Any probability density function $p(y|\theta)$ can be written as

$$p(y|\theta) = \frac{q_\theta(y)}{Z(\theta)} \quad (11)$$

where $q_\theta(y)$ is the unnormalised density and the normalising constant $Z(\theta) = \int p(y, \theta) dy$ is unknown. In this case the acceptance ratio of the Metropolis-Hastings algorithm is

$$\alpha = \min \left(1, \frac{q_{\theta^*}(y)\pi(\theta^*)Z(\theta)h(\theta|\theta^*)}{q_\theta(y)\pi(\theta)Z(\theta^*)h(\theta^*|\theta)} \right) \quad (12)$$

where $\pi(\theta)$ is the prior distribution of θ . The acceptance ratio in (12) involves computing unknown normalising constants. Introducing auxiliary variables θ^*, y^* and sampling from an augmented distribution

$$\pi(\theta^*, y^*, \theta|y) \propto p(y|\theta)\pi(\theta)p(y^*|\theta^*)h(\theta^*|\theta) \quad (13)$$

results in

$$\alpha = \min \left(1, \frac{p(y|\theta^*)\pi(\theta^*)p(y^*|\theta)h(\theta|\theta^*)}{p(y|\theta)\pi(\theta)p(y^*|\theta^*)h(\theta^*|\theta)} \right) \quad (14)$$

$$= \min \left(1, \frac{q_\theta(y^*)\pi(\theta^*)h(\theta|\theta^*)q_{\theta^*}(y)Z(\theta)Z(\theta^*)}{q_\theta(y)\pi(\theta)h(\theta^*|\theta)q_{\theta^*}(y^*)Z(\theta^*)Z(\theta)} \right) \quad (15)$$

$$= \min \left(1, \frac{q_\theta(y^*)\pi(\theta^*)q_{\theta^*}(y)}{q_\theta(y)\pi(\theta)q_{\theta^*}(y^*)} \right) \quad (16)$$

where the normalising constants cancel out and $h(\cdot)$ is a symmetric distribution [9, 8]. In order to be able to use this algorithm one has to be able to sample from the unnormalised density which in the case of the COM-Poisson distribution can be done efficiently using rejection sampling.

Updating the parameter μ of the COM-Poisson we have $\theta = (\mu, \nu)$ and $\theta^* = (\mu^*, \nu)$ where μ^* follows a Normal distribution centered at μ and

$$q_\theta(y^*) = \left(\frac{\mu_i^{y_i^*}}{y_i^*!} \right)^{\nu_i} \quad q_{\theta^*}(y) = \left(\frac{(\mu_i^*)^{y_i}}{y_i!} \right)^{\nu_i} \quad (17)$$

$$q_\theta(y) = \left(\frac{\mu_i^{y_i}}{y_i!} \right)^{\nu_i} \quad q_{\theta^*}(y^*) = \left(\frac{(\mu_i^*)^{y_i^*}}{y_i^*!} \right)^{\nu_i} \quad (18)$$

Likewise for updating the parameter ν .

3 Bayesian density regression

Density regression is similar to quantile regression in that it allows flexible modelling of the response variable Y given the covariates $\mathbf{x} = (x_1, \dots, x_p)'$. Features (mean, quantiles, spread) of the conditional distribution of the response variable, vary with \mathbf{x} , so, depending on the predictor values, features of the conditional distribution can change in a different way than the population mean. The difference between density regression and quantile regression is that density regression models the probability density function or probability mass function rather than directly modelling the quantiles.

3.1 Bayesian density regression for count data

This paper focuses on the following mixture of regression models:

$$f(y_i|\mathbf{x}_i) = \int f(y_i|\mathbf{x}_i, \phi_i) G_{\mathbf{x}_i}(\mathrm{d}\phi_i) \quad (19)$$

where

$$f(y_i|\mathbf{x}_i, \phi_i) = \text{COM-P}(y_i; \exp(\mathbf{x}_i' \mathbf{b}_i), \exp(\mathbf{x}_i' \mathbf{c}_i)) \quad (20)$$

the conditional density of the response variable given the covariates is expressed as a mixture of COM-Poisson regression models with $\phi_i = (\mathbf{b}_i, \mathbf{c}_i)$ and $G_{\mathbf{x}_i}$ is an unknown mixture distribution that changes according to the location of \mathbf{x}_i .

3.2 MCMC algorithm

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ denote the $k \leq n$ distinct values of ϕ and let $\mathbf{S} = (S_1, \dots, S_n)'$ be a vector of indicators denoting the global configuration of subjects to distinct values $\boldsymbol{\theta}$, with $S_i = h$ indexing the location of the i th subject within the $\boldsymbol{\theta}$. In addition, let $\mathbf{C} = (C_1, \dots, C_k)'$ with $C_h = j$ denoting that θ_h is an atom from the basis distribution, $G_{\mathbf{x}_j}^*$. Hence $C_{S_i} = Z_i = j$ denotes that subject i is drawn from the j th basis distribution.

Excluding the i th subject, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta} \setminus \{\phi_i\}$ denotes the $k^{(i)}$ distinct values of $\phi^{(i)} = \phi \setminus \{\phi_i\}$, $\mathbf{S}^{(i)}$ denotes the configuration of subjects $\{1, \dots, n\} \setminus \{i\}$ to these values and $\mathbf{C}^{(i)}$ indexes the DP component numbers for the elements of $\boldsymbol{\theta}^{(i)}$.

Grouping the subjects in the same cluster and updating the prior with the likelihood for the data \mathbf{y} , we obtain the conditional posterior

$$(\phi_i | \mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \mathbf{X}, a) \sim q_{i0} G_{i,0} + \sum_{h=1}^{k^{(i)}} q_{ih} \delta_{\theta_h^{(i)}}, \quad (21)$$

where $G_{i,0}(\phi)$ is the posterior obtained by updating the prior $G_0(\phi)$ and the likelihood $f(y_i | \mathbf{x}_i, \phi)$:

$$G_{i,0}(\phi) = \frac{G_0(\phi) f(y_i | \mathbf{x}_i, \phi)}{h_i(y_i | \mathbf{x}_i)}, \quad (22)$$

$$q_{i0} = c w_{i0} h_i(y_i | \mathbf{x}_i), \quad q_{ih} = c w_{ih} f(y_i | \mathbf{x}_i, \theta_h), \quad (23)$$

$$w_{i0} = \sum_{j=1}^n \frac{a b_{ij}}{a + \sum_{l \neq i} \mathbf{1}(C_{S_l^{(i)}}^{(i)} = j)}, \quad w_{ih} = \frac{b_{i, C_h^{(i)}} \sum_{m \neq i} \mathbf{1}(S_m^{(i)} = h)}{a + \sum_{l \neq i} \mathbf{1}(C_{S_l^{(i)}}^{(i)} = C_h)}, \quad (24)$$

where b_{ij} are weights that depend on the distance between subjects' predictor values, c is a normalising constant and $h = 1, \dots, k^{(i)}$. Since there is no closed form expression for the posterior distribution, approximation of the probability $q_{i0} = c w_{i0} h_i(y_i | \mathbf{x}_i)$ is difficult.

We overcome this problem by bridging: i) an MCMC algorithm for sampling from the posterior distribution of a Dirichlet process model, with a non-conjugate prior, found in [10]; ii) the MCMC algorithm found in [3]; and iii) a variation of the MCMC exchange algorithm.

The MCMC algorithm alternates between the following steps:

Step 1: Update S_i for $i = 1, \dots, n$, by proposing, from the conditional prior, a move to a new cluster or an already existing cluster with probabilities proportional to w_{i0} and w_{ih} for $h = 1, \dots, k^{(i)}$.

- a) If the proposed move is to go to a new cluster we draw parameters (μ_0, ν_0) for that cluster from G_0 and at the same time sample an observation y^* from the COM-Poisson(μ_0, ν_0). The acceptance ratio of the Metropolis-Hastings algorithm is

$$\min \left(1, \frac{q_{\theta}(y^*) q_{\theta^*}(y)}{q_{\theta}(y) q_{\theta^*}(y^*)} \right) \quad (25)$$

If the proposal is accepted, $C_{S_i} \sim \text{multinomial}(\{1, \dots, n\}, \mathbf{b}_i)$.

- b) If the proposed move is to an already existing cluster h , we sample an observation y^* from the COM-Poisson(μ_h, ν_h) and accept with the same probability as in (25). If the proposal is accepted $C_{S_i} = C_h$.

Step 2: Update the parameters θ_h , for $h = 1, \dots, k$ by sampling from the conditional posterior distribution

$$(\theta_h | \mathbf{S}, \mathbf{C}, \boldsymbol{\theta}^{(h)}, k, \mathbf{y}, \mathbf{X}) \sim \prod_{i:S_i=h} f(y_i | \mathbf{x}_i, \theta_h) \} G_0(\theta_h), \quad (26)$$

using the Metropolis-Hasting algorithm with acceptance probability as in (16).

Step 3: Update C_h , for $h = 1, \dots, k$, by sampling from the multinomial conditional with

$$(C_h | \mathbf{S}, \mathbf{C}^{(h)}, \boldsymbol{\theta}, k, \mathbf{y}, \mathbf{X}) \sim \frac{\prod_{i:S_i=h} b_{ij}}{\sum_{l=1}^n \prod_{i:S_i=h} b_{il}}, \quad j = 1, \dots, n \quad (27)$$

and location weights γ_j for $j = 1, 2, \dots, n$, using an approach used in [4].

4 Simulations and application

We consider two simulated data sets to compare the proposed discrete Bayesian density regression method to the “jittering” method. These are

$$Y_i | X_i = x_i \sim \text{Binomial}(10, 0.3x_i) \quad (28)$$

$$Y_i | X_i = x_i \sim 0.4\text{Pois}(\exp(1 + x_i)) + 0.2\text{Binomial}(10, 1 - x_i) + 0.4\text{Geom}(0.2) \quad (29)$$

where $x_i \sim \text{Unif}(0, 1)$. Table (1) shows the absolute mean errors obtained using both methods. If q_p is the true conditional quantile when $x = p$ and \hat{q}_p is the estimated conditional quantile, the mean absolute error is defined as $\mathbb{E}[|q_p - \hat{q}_p|]$. The discrete Bayesian density regression (BDR) estimates outperform the “jittering” method and in almost all cases the “jittering” method leads to crossing quantiles (except when $n = 500$).

Method	Number of Observations					
	Binomial			Mixture		
	20	100	500	20	100	500
Density Regression	0.5576	0.2820	0.2421	0.7435	0.5833	0.3589
Jittering (linear)	0.5256	0.8461	0.4765	1.1923	0.6666	0.4294
Jittering (splines)	0.7820	0.5128	0.3020	1.9487	0.8269	0.3910

Table 1: Mean absolute error obtained using the different density/quantile regression methods.

We apply the discrete density regression technique to data on housebreakings in Greater Glasgow (Scotland). The data consist of the number of housebreakings in each of the 127 intermediate geographies in Greater Glasgow in 2010. We aim to relate the number of housebreakings to the deprivation score of the intermediate geography area, as measured by the Scottish Index of Multiple Deprivation (SIMD). The deprivation score is standardised by considering the difference of each intermediate geography’s deprivation from the average deprivation in Greater Glasgow e.g. low values relate to affluent areas, large values to deprived areas. The solid and dashed lines in figure 1 show the quantiles (for $p = 0.1, 0.5, 0.95$) obtained for the standard Poisson regression model and the COM-Poisson model respectively. The first model is not able to capture the overdispersion of the data, nor the skewness of the distribution.

5 Conclusions and further research

In this manuscript we have proposed a novel Bayesian density regression technique for discrete data which is based on mixing COM-Poisson distributions. The new method takes advantage of the exchange algorithm and updates the cluster allocations by drawing a new allocation for an auxiliary observation and then accepting or rejecting it. As a result the MCMC samples from the target distribution without the need to estimate the normalisation constant of the likelihood. The method overcomes the two main drawbacks of the “jittering” method for discrete quantile regression, namely that it does not require the addition of artificial additional noise and that

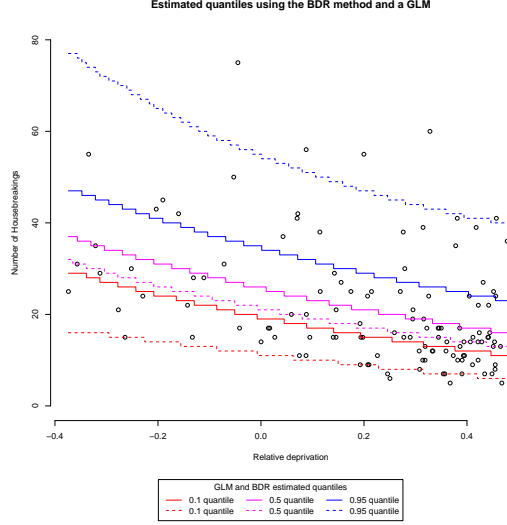


Figure 1: Estimated quantiles for housebreaking data, using discrete Bayesian density regression (dashed lines) and derived from a Poisson model.

it does not suffer from the problem of crossing quantiles. We have illustrated the method in a real world application as well as simulated examples in which our method compared favourably to the “jittering” method. Further research efforts will be devoted in improving the computational speed and efficiency of the MCMC algorithm to make it an even more attractive alternative to “jittering”.

References

- [1] Buchinsky, Moshe (1998) *Recent advances in quantile regression models: A practical guideline for empirical research*. The journal of human resources, **33**, 88–126.
- [2] Conway, Richard W. and Maxwell, William L. (1962) *A queuing model with state dependent service rate*. Journal of industrial engineering, **12**, 132–136.
- [3] Dunson, David B. and Pillai, Natesh and Park, Ju-Hyun. (2007) *Bayesian density regression*. Journal of the royal statistical society: Series B, **69**, 163–183.

- [4] Dunson, David B. and Stanford, Joseph B. (2005) *Bayesian inferences on predictors of conception probabilities*. Biometrics, **61**, 126–133.
- [5] Karabatsos, George and Walker, Stephen G. (2012) *Adaptive-modal Bayesian nonparametric regression*. Electronic journal of statistics, 6, 2038–2068.
- [6] Koenker, Roger and Bassett, Gilbert (1978) *Regression quantiles*. Econometrica, **46**, 33–50.
- [7] Machado, Josè António Ferreira and Santos Silva, João M.C.. (2005) *Quantiles for counts*. Journal of the american statistical association, **100**, 1226–1237.
- [8] Møller, J. and Pettitt, A. N. and Reeves, R. and Berthelsen, K. K. (2006) *An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants*. Biometrika, **932**, 451–458.
- [9] Murray, Ian and Ghahramani, Zoubin and MacKay, David J. C. (2006) *MCMC for doubly-intractable distributions*. Proceedings of the 22nd Annual UAI Conference, 359–366.
- [10] Neal, Radford M. (2000) *Markov chain sampling methods for Dirichlet process mixture models*. Journal of computational and graphical statistics, **9**, 249–265.
- [11] Shmueli, Galit and Minka, Thomas P. and Kadane, Joseph B. and Borle, Sharad and Boatwright, Peter (2008) *A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution*. Journal of the royal statistical society: Series C, **54**, 127–142.